Synthese (2016) 193:1011–1027
DOI 10.1007/s11229-015-0673-9

The three faces of faithfulness

Jiji Zhang · Peter Spirtes

Received: 22 April 2014 / Accepted: 13 January 2015 / Published online: 11 February 2015

Abstract In the causal inference framework of Spirtes, Glymour, and Scheines (SGS), inferences about causal relationships are made from samples from probability distributions and a number of assumptions relating causal relations to probability distributions. The most controversial of these assumptions is the Causal Faithfulness Assumption, which roughly states that if a conditional independence statement is true of a probability distribution generated by a causal structure, it is entailed by the causal structure and not just for particular parameter values. In this paper we show that the addition of the Causal Faithfulness Assumption plays three quite different roles in the SGS framework: (i) it reduces the degree of underdetermination of causal structure by probability distribution; (ii) computationally, it justifies reliable (constraint-based) causal inference algorithms that would otherwise have to be slower in order to be reliable; and (iii) statistically, it implies that those algorithms reliably obtain the correct answer at smaller sample sizes than would otherwise be the case. We also consider a number of variations on the Causal Faithfulness Assumption, and show how they affect each of these three roles.

Keywords Causal inference · Bayes nets · Faithfulness · Graphical models

J. Zhang
Department of Philosophy, Lingnan University, Room HSH201, Ho Sin Hang Building, Tuen Mun, N.T., Hong Kong
e-mail: jijizhang@ln.edu.hk

P. Spirtes (✉)
Department of Philosophy, Carnegie Mellon University, 135D Baker Hall, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
e-mail: ps7z@andrew.cmu.edu

1 Introduction

In the causal inference framework of Spirtes, Glymour, and Scheines (SGS, 2000), inferences about causal structures are made from patterns of conditional independence and dependence relations that are estimated from samples. The inference procedures are justified by, among other things, assumptions relating causal structures to probability distributions. Among these assumptions the two best known are the Causal Markov Assumption (CMA) and the Causal Faithfulness Assumption (CFA). Roughly, the CMA states that the causal structure of a suitably chosen set of variables entails a set of conditional independence statements that must be satisfied by the joint probability distribution of these variables, and the CFA states that the joint probability distribution satisfy *only* those conditional independence statements that are entailed by the causal structure according to the CMA.

Both assumptions have been occasions of debates (e.g., Woodward 1998; Hausman and Woodward 1999; Cartwright 2001; Hoover 2001; Steel 2006; Andersen 2013), but the CFA is generally regarded as more questionable and is often defended as a methodological assumption in the philosophical literature (Hitchcock 2010). In this paper we will not enter the debate on the merits of the assumption (though what we have to say will be relevant to that debate), but aim to clarify the consequences of adding the CFA by distinguishing three roles it plays in the SGS framework: first, it serves to reduce underdetermination of causal structure by probability distribution; second, computationally it justifies reliable (constraint-based) causal inference algorithms that would otherwise have to be slower; third, statistically it allows those algorithms to work at smaller sample sizes than would otherwise be the case. Our discussion will draw on recent results on a number of variations on the CFA, and we show how they affect each of these three roles.

We will proceed as follows. In Sect. 2, we review the basics of the SGS framework of causal inference. In Sect. 3, we examine the role of reducing underdetermination of causal structure by probability distribution, and show that in a way some weaker versions of the CFA play this role even better. However, superiority in playing the first role comes with inferiority in playing the second and the third roles, as we shall explain in Sects. 4 and 5. We conclude in Sect. 6.

We will not argue that the CFA can be justified by appealing to its computational and statistical consequences, and it is not our present purpose to defend the CFA against the weaker assumptions. Our intention is rather to clarify the prices one may have to pay to adopt the weaker assumptions. Whether or not they are worth paying is a separate issue and is in all likelihood context dependent.

2 Preliminaries

In the SGS framework, a causal system is characterized by a set of random variables $\mathbf{V}$, and the causal structure over $\mathbf{V}$ is understood as the set of *direct causal relations* between variables relative to $\mathbf{V}$. It is convenient to represent the causal structure by a directed graph: variables in $\mathbf{V}$ are represented by distinct nodes in the graph, and a directed edge or arrow ($\rightarrow$) is drawn from the node representing variable X to the

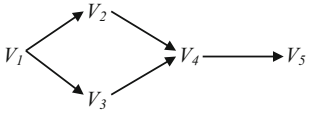

Fig. 1 An acyclic causal structure over five variables

node representing variable Y if and only if X has a direct causal influence on Y relative to $\mathbf{V}$. We call this representation a *causal graph*, and we assume that direct causation between variables is irreflexive and antisymmetric, so that in a causal graph there is no arrow from a node to itself and there is at most one arrow between any two nodes. For convenience, we use "causal structure" and "causal graph" interchangeably (as well as "variables" and "nodes"), as their differences do not matter for our purposes.

Some graph terminology will be helpful. In a directed graph, nodes X and Y are *adjacent* if there is an arrow between them in either direction. If the arrow is $X \rightarrow Y$, X is called a *parent* of Y and Y a *child* of X. A *directed path* is an ordered sequence of two or more distinct nodes such that every node except for the last one in the sequence is a parent of its successor in the sequence. X is called an *ancestor* of Y and Y a *descendant* of X if $X = Y$ or there is a directed path from X to Y.[1] A *directed cycle* occurs in the graph if there are two distinct nodes that are ancestors of each other. A directed graph is called *acyclic* if there is no directed cycle in the graph. For example, the graph in Fig. 1 is acyclic. In the graph, V_1 has no parent, and every node is its descendant. V_2 has a single parent V_1, and has three descendants $\{V_2, V_4, V_5\}$. Similarly, V_3 has a single parent V_1, and has three descendants $\{V_3, V_4, V_5\}$. V_4 has two parents $\{V_2, V_3\}$, and two descendants $\{V_4, V_5\}$. Finally, V_5 has one parent V_4, and one descendant V_5.

In this paper we consider only causal structures that can be properly represented by directed acyclic graphs (DAGs). Given a causal DAG G and a joint probability distribution P over a set of variables $\mathbf{V}$, G and P satisfy the (local) *Markov property*— in which case we also say P is Markov to G and G is Markov to P— if and only if according to P, every variable is conditionally independent of the set of variables that are neither its descendants nor its parents in G given the set of its parents in G. For example, if G is the DAG in Fig. 1, then according to the Markov property, each of the five variables is independent of its non-descendants given its parents [2]:

(1) $V_1 \perp\!\!\!\perp \varnothing | \varnothing$, i.e., V_1 is independent of the empty set conditional on the empty set.
(2) $V_2 \perp\!\!\!\perp V_3 | V_1$, i.e., V_2 is independent of V_3conditional on V_1.
(3) $V_3 \perp\!\!\!\perp V_2 | V_1$, i.e., V_3 is independent of V_2 conditional on V_1.
(4) $V_4 \perp\!\!\!\perp V_1 | \{V_2, V_3\}$, i.e., V_4 is independent of V_1 conditional on $\{V_2, V_3\}$.
(5) $V_5 \perp\!\!\!\perp \{V_1, V_2, V_3\} | V_4$, i.e., V_5 is independent of $\{V_1, V_2, V_3\}$ conditional on V_4.

[1] The stipulation that a node counts as its own ancestor and descendant has some technical convenience.

[2] For three disjoint sets of random variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, we write $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$ for the statement that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$. The statement is satisfied by a probability distribution P or true according to P iff $P(\mathbf{X} = x | \mathbf{Y} = y, \mathbf{Z} = z) = P(\mathbf{X} = x | \mathbf{Z} = z)$ for every (vector) value x of $\mathbf{X}$, y of $\mathbf{Y}$, z of $\mathbf{Z}$ s.t. $P(\mathbf{Y} = y, \mathbf{Z} = z) > 0$. When $\mathbf{Y}$ is an empty set, $\mathbf{X} \perp\!\!\!\perp \varnothing | \mathbf{Z}$ is trivially satisfied by all probability distributions. When $\mathbf{Z}$ is empty, we often just write $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$, short for $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \varnothing$. For singleton sets, we will abuse notation and omit the curly brackets. See e.g., Dawid (1979) and Pearl (1988) for detailed discussions of conditional independence.

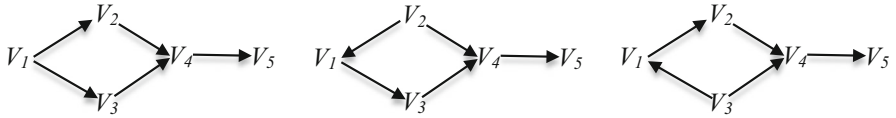

Fig. 2 Causal structures that are Markov equivalent to the structure in Fig. 1

These five conditional independence statements are what the (local) Markov property explicitly requires. Not every one of the five is nontrivial: every probability distribution trivially satisfies (1). Some DAGs, i.e. complete DAGs in which every two nodes are adjacent, do not entail any nontrivial conditional independence statement by the Markov property. Moreover, the nontrivial conditional independence statements explicitly required by the Markov property, such as (5), may entail other nontrivial conditional independence statements by the axioms of probability calculus, such as $V_5 \perp\!\!\!\perp V_1 | V_4$. We will refer to all these nontrivial conditional independence statements that follow from the Markov property as conditional independence statements *entailed* by the DAG.[3]

The CMA states that the true causal structure of **V** and the true probability distribution of **V** satisfy the Markov property. Throughout the paper we assume that we are working with a suitably chosen set of variables **V** that satisfies the CMA.[4]

Under the CMA, some hypotheses of the causal structure can be refuted by a probability distribution. But the assumption by itself is unable to rule out enough hypotheses to justify interesting causal inference. For example, as already mentioned, a complete DAG does not entail any nontrivial conditional independence statement, and so no structure represented by a complete DAG is falsifiable by any distribution given only the CMA. Then every causal arrow remains a possibility because it appears in some complete structure.

This radical underdetermination is significantly reduced when we add the CFA. A probability distribution P is *faithful* to a DAG G if and only if all conditional independence statements satisfied by P are entailed by G. The CFA states that the true distribution is faithful to the true causal structure. Under this assumption, a complete structure, for example, is refuted whenever a nontrivial conditional independence statement is true of the given distribution.

Under the CMA and the CFA, for many distributions, the underdetermination of causal structure is sufficiently reduced to allow informative causal inference. For example, suppose we are given a distribution that is both Markov and faithful to the DAG in Fig. 1. Then under the two assumptions, the only (acyclic) causal structures over the five variables that are compatible with the distribution are (a), (b), and (c) in Fig. 2. These candidates share interesting structural features, including adjacencies and most conspicuously, the arrows $V_2 \rightarrow V_4$, $V_3 \rightarrow V_4$, and $V_4 \rightarrow V_5$.

[3] All the entailed conditional independence statements can be easily read off a given causal structure by a graph criterion called d-separation (Pearl 1988).

[4] This will generally be the case if **V** does not leave out any common causes of two variables in **V**, there is no selection bias in the sampling, and the true causal structure is either acyclic or linear.

Computationally, various algorithms have been developed to search for these compatible causal structures given a distribution (or more realistically, samples from a distribution). One approach, known as the constraint-based approach and championed by SGS, is to systematically recover features of the underlying causal structure from conditional independence facts that are (judged to be) true of the distribution.

Recently, it has been shown that given the assumption of no unmeasured common causes together with a variety of distributional assumptions, [e.g. non-Gaussian linear models (Shimizu et al. 2006), or models with additive noise (Peters et al. 2011; Hoyer et al. 2008)], but without assuming the CFA, there are causal inference algorithms that can both reduce the amount of underdetermination of causal models from data and improve the accuracy of the output, as compared to constraint-based algorithms. In cases where the distributional assumptions and the causal assumptions are warranted there is good reason to apply these alternative algorithms in place of constraint-based algorithms. However, there are a number of advantages of constraint-based algorithms that still make them useful for certain domains of applications.

(1) Constraint-based algorithms apply quite generally. As long as consistent tests of conditional independence are available, constraint-based algorithms can be reliably (in the large sample limit) applied. Distribution free tests of conditional independence using kernel methods (Zhang et al. 2011) have recently been developed (although they typically require large sample sizes to be reliable and are computationally slow).
(2) For the Gaussian and multinomial distributions we are not aware of any alternatives to constraint-based algorithms that can handle as many variables in a feasible amount of time. This is important for cases such as inferring genetic regulatory networks, where the number of variables can easily run into the many thousands.
(3) There are extensions of constraint-based methods to the case where there may be hidden common causes and selection bias (e.g. the RFCI and the FCI+ algorithms), which also can run on hundreds of variables (in the Gaussian or multinomial case), and on smaller sets of variables where further distributional assumptions are not warranted. Although there have been a few attempts to extend some alternatives to constraint-based algorithms to cases where there may be unmeasured common causes (e.g. in the case of linear non-Gaussian models) they have required large sample sizes and cannot be applied to large numbers of variables.
(4) Some of the distributional assumptions that are made for the case of no unmeasured common causes (e.g., additive noise models) are not preserved under marginalization of hidden variables.

Our subsequent discussions will focus on the constraint-based approach, which exploits, among other things, two facts:

Proposition 1 (Spirtes et al. 2000, p. 47)*: Two variables are adjacent in a DAG if and only if they are not entailed to be independent conditional on any subset of other variables in the DAG.*

The other fact has to do with *unshielded triples* in a DAG. An ordered triple of variables $<X, Y, Z>$ is *unshielded* in a DAG if X and Z are both adjacent to Y, but

X and Z are not adjacent. The triple is called an *unshielded collider* if the two arrows both point to Y (i.e., $X \rightarrow Y \leftarrow Z$); otherwise it is an *unshielded noncollider*.

Proposition 2 (Spirtes et al. 2000, p. 47)*: Let* $<X, Y, Z>$ *be any unshielded triple in a DAG. Then*

(1) $<X, Y, Z>$ *is an unshielded collider if and only if* X *and* Z *are not entailed to be independent conditional on any subset of other variables that contains* Y.
(2) $<X, Y, Z>$ *is an unshielded non-collider if and only if* X *and* Z *are not entailed to be independent conditional on any subset of other variables that does not contain* Y.

These two facts immediately suggest a causal discovery procedure known as the SGS algorithm (Spirtes et al. 2000, p. 82). The algorithm, however, is computationally and statistically inefficient. Fortunately, the CFA also justifies more efficient procedures such as the well-known PC algorithm. Thus the CFA plays a role in boosting computational efficiency and statistical efficiency, besides its role of reducing underdetermination of causal structure by probability distribution. These roles should be distinguished, for some weaker assumptions of faithfulness play the role of reducing underdetermination even better, but they do not play the other roles as well.

3 The role of reducing underdetermination of causal structure by probability distribution

For any probability distribution P of $\mathbf{V}$, let $\mathbf{M}(P)$ denote the set of causal structures (which, recall, are restricted to DAGs) over $\mathbf{V}$ to which P is Markov, and $\mathbf{M}\text{-}\mathbf{F}(P)$ denote the set of causal structures to which P is both Markov and faithful. Given a distribution P, if we just assume the CMA, the set of causal structures that is underdetermined by P is $\mathbf{M}(P)$. If we further add the CFA, the set of causal structures that is underdetermined by P is $\mathbf{M}\text{-}\mathbf{F}(P)$. $\mathbf{M}\text{-}\mathbf{F}(P)$ is a subset of $\mathbf{M}(P)$ for every P, and is a proper subset unless P does not satisfy any nontrivial conditional independence statement.[5] This is the sense in which adding the CFA reduces underdetermination of causal structure by a probability distribution.

There is a catch. Although $\mathbf{M}(P)$ is always non-empty, $\mathbf{M}\text{-}\mathbf{F}(P)$ is empty for some P. That is, some distributions are not Markov and faithful to any causal structure (Zhang and Spirtes 2008). Consider the following:

Example 3.1 Among the distributions that are Markov to the causal structure in Fig. 1, some can satisfy one (and only one) extra independence statement: $V_1 \perp\!\!\!\perp V_4$, due perhaps to an exact balancing out of the two causal paths, $V_1 \rightarrow V_2 \rightarrow V_4$ and $V_1 \rightarrow V_3 \rightarrow V_4$. Such a distribution is not only unfaithful to the structure in Fig. 1, it is not both Markov and faithful to any DAG over the five variables. (Recall that we consider only DAG structures in this paper).

[5] If P does not satisfy any nontrivial conditional independence statement, then $\mathbf{M}(P) = \mathbf{M}\text{-}\mathbf{F}(P)$ is the set of complete causal structures.

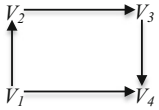

Fig. 3 An example that violates Adjacency-Faithfulness (Example 3.2)

When **M-F**(P) is empty, it is obviously inappropriate to say that the underdetermination of causal structure by P is reduced by adding the CFA. Rather, under the supposition of the CMA, the emptiness of **M-F**(P) means that the CFA cannot be true for that probability distribution, or say, in a methodological description, that the CFA does not apply to that probability distribution.

So the overall picture is that the CFA applies to some but not all distributions. For those it applies to, that is, for those P such that **M-F**(P) is non-empty, adopting the CFA in addition to the CMA reduces underdetermination of causal structure from **M**(P) to **M-F**(P).

From this perspective, several weaker variations on the CFA actually play the role of reducing underdetermination even better: they apply to more distributions than the CFA does, and reduce underdetermination as much as the CFA does when the latter applies.

Let us consider an increasingly weaker sequence of variations. The first is named the *Adjacency-Faithfulness* assumption (Ramsey et al. 2006; Zhang and Spirtes 2008).[6] It states that if two variables are adjacent in the true causal structure, then they are not independent conditional on any subset of other variables. This assumption follows from the CFA (in view of Proposition 1), and is strictly weaker. For example, the case in Example 3.1 does not satisfy the CFA but satisfies the Adjacency-Faithfulness assumption. Since the distribution in Example 3.1 is one to which the CFA does not apply, the Adjacency-Faithfulness assumption applies to more distributions than the CFA does.[7]

There are, however, also probability distributions that are both Markov and Adjacency-Faithful to no causal structure.

Example 3.2 Consider the causal structure in Fig. 3. Among the distributions that are Markov to this structure, some satisfy one and only one extra conditional independence statement $V_1 \perp\!\!\!\perp V_4$, due for example to an exact balancing out of the two causal paths, $V_1 \rightarrow V_2 \rightarrow V_3 \rightarrow V_4$ and $V_1 \rightarrow V_4$. Such a distribution is not both Markov and Adjacency-Faithful to any DAG over the four variables.[8]

[6] Another consequence of the CFA, named *Orientation-Faithfulness*, is often discussed alongside Adjacency-Faithfulness. The Adjacency-Faithfulness assumption, as we will show, can reduce underdetermination as much as the CFA does without Orientation-Faithfulness, but the Orientation-Faithfulness assumption cannot do without Adjacency-Faithfulness.

[7] Interested readers can check that given the CMA and the Adjacency-Faithfulness assumption, the causal structures that are compatible with the said distribution remain the three structures in Fig. 2.

[8] Basically the reason is that no DAG with just three adjacencies (one between V_1 and V_2, one between V_2 and V_3, and one between V_3 and V_4) is Markov to the said distribution. Adding any other adjacency yields a DAG to which the said distribution is not Adjacency-Faithful.

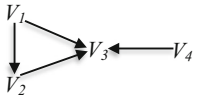

Fig. 4 An example that violates Triangle-Faithfulness (Example 3.3)

Hence the Adjacency-Faithfulness assumption does not apply to all distributions, and it is worth considering even weaker assumptions. The next in line is a conjunction of two assumptions, both of which are entailed by the Adjacency-Faithfulness assumption. One is known as the causal Minimality assumption (Spirtes et al. 2000, p. 31), which states that no proper subgraph of the true causal DAG is Markov to the true probability distribution. Suppose the distribution is Adjacency-Faithful to the causal structure. Taking away any arrow, say $X \rightarrow Y$, would yield a structure that entails a conditional independence between X and Y, but as a logical consequence of Adjacency-Faithfulness, no statement of conditional independence between X and Y is true of the distribution. So the resulting substructure is not Markov to the distribution.

The other is named the *Triangle-Faithfulness* assumption (Zhang and Spirtes 2008): Let X, Y, Z be any three variables that form a triangle in the causal structure of $\mathbf{V}$ (i.e., they are adjacent to one another):

(1) If Z is a non-collider on the path $<X, Z, Y>$, then X and Y are not independent conditional on any subset of $\mathbf{V}\backslash\{X, Y\}$ that does not contain Z;
(2) If Z is a collider on the path $<X, Z, Y>$, then X and Y are not independent conditional on any subset of $\mathbf{V}\backslash\{X, Y\}$ that contains Z.

Clearly the Triangle-Faithfulness assumption is also a logical consequence of the Adjacency-Faithfulness assumption. Thus the conjunction of the causal Minimality assumption and the Triangle-Faithfulness assumption is entailed by the Adjacency-Faithfulness assumption. The former is strictly weaker. For example, the case in Example 3.2 violates the Adjacency-Faithfulness assumption, but satisfies the causal Minimality assumption and the Triangle-Faithfulness assumption.

However, the weaker conjunction still does not apply to all distributions. An example:

Example 3.3 Among the distributions that are Markov to the causal structure depicted in Fig. 4, some imply one (and only one) extra conditional independence statement: $V_1 \perp\!\!\!\perp V_2 | V_3$. Such a distribution is then not Triangle-Faithful to the structure in Fig. 4. Moreover, it is not both Markov and Triangle-Faithful to any DAG over the four variables.[9]

Finally, there is another minimality assumption formulated in Pearl (2009) that is closely related to the CFA. Following Zhang (2013), we will refer to it as the *P-Minimality* (Pearl's minimality) assumption. For any two DAGs $G$and H over $\mathbf{V}$, call

[9] Basically the reason is that every DAG to which the said distribution is Markov and Triangle-Faithful must have at least the three arrows $V_1 \rightarrow V_3$, $V_2 \rightarrow V_3$, and $V_3 \rightarrow V_4$ (cf. the VCSGS algorithm we will describe in the next section). Since it is not true that $V_1 \perp\!\!\!\perp V_2$, an edge between V_1 and V_2 has to be added in order to be Markov to the distribution. Then the triangle $<V_1, V_3, V_2>$ would fail the Triangle-Faithfulness.

Table 1 Examples that illustrate the applicability of the sequence of assumptions

	Example 3.1	Example 3.2	Example 3.3
M-F(P)	$= \varnothing$	$= \varnothing$	$= \varnothing$
M-AF(P)	$\neq \varnothing$	$= \varnothing$	$= \varnothing$
M-M-TF(P)	$\neq \varnothing$	$\neq \varnothing$	$= \varnothing$
M-PM(P)	$\neq \varnothing$	$\neq \varnothing$	$\neq \varnothing$

H an *I-submodel* (Independence-submodel)[10] of G if every conditional independence statement entailed by G is also entailed by H; H is a *proper* I-submodel of Gif H is an I-submodel of G but not vice versa. The P-Minimality assumption states that no proper I-submodel of the true causal DAG is Markov to the true probability distribution.

As shown in Zhang (2013), the conjunction of the causal Minimality assumption and the Triangle-Faithfulness assumption entails the P-Minimality assumption, but not vice versa. For example, the case in Example 3.3 violates the Triangle-Faithfulness assumption, but satisfies the P-Minimality assumption.

It is also easy to see that the P-Minimality assumption applies to all distributions, for given any distribution, there is always a minimal DAG in the I-submodel ordering that is Markov to the distribution.

Let **M-AF**(P) denote the causal structures that are compatible with P under the causal Markov and Adjacency-Faithfulness assumptions, **M-M-TF**(P) denote the set of causal structures that are compatible with P under the causal Markov, Minimality, and Triangle-Faithfulness assumptions, and **M-PM**(P) denote the set of causal structures that are compatible with P under the causal Markov and P-Minimality assumptions. What we have said so far implies that for every P,

$$\textbf{M-F}\,(P) \subseteq \textbf{M-AF}\,(P) \subseteq \textbf{M-M-TF}\,(P) \subseteq \textbf{M-PM}\,(P)\,,$$

for the sequence of assumptions is increasingly weaker. Moreover, for some P, **M-F**(P) is empty but **M-AF**(P) is not; for some P, **M-AF**(P) is empty but **M-M-TF**(P) is not; and for some P, **M-M-TF**(P) is empty but **M-PM**(P) is not. In other words, the increasingly weaker assumptions apply to increasingly more distributions, with the weakest P-Minimality assumption applying to all. (See Table 1 for a summary with respect to the distributions described in the previous examples.)

For example, for the distribution in Example 3.1, the set of causal structures that are Markov and Faithful to the distribution is empty, but the set of causal structures that are Markov and Adjacency-Faithful to the distribution is not.

We now come to the main point. We will show that whenever **M-F**(P) is non-empty,

$$\textbf{M-F}\,(P) = \textbf{M-AF}\,(P) = \textbf{M-M-TF}\,(P) = \textbf{M-PM}\,(P)\,.$$

[10] It is a *sub*-model in that it entails a superset of conditional independence constraints and so is compatible with a *sub*set of probability distributions.

Given the subset relationship established above, it suffices to show that if $\mathbf{M}\text{-}\mathbf{F}(P) \neq \varnothing$, then $\mathbf{M}\text{-}\mathbf{PM}(P) \subseteq \mathbf{M}\text{-}\mathbf{F}(P)$. In words, this says that if there is a causal structure to which P is both Markov and faithful, then for every structure G, G and P satisfy the Markov and P-Minimality assumptions only if G and P satisfy the Markov and Faithfulness assumptions.

This is fairly easy to prove. Suppose G and P satisfy the Markov and P-Minimality assumptions. Let H be one of the causal structures to which P is both Markov and faithful. Since G is Markov to P, all conditional independence statements entailed by G are satisfied by P. Since P is faithful to H, all conditional independence statements satisfied by P are entailed by H. Thus, all conditional independence statements entailed by G are entailed by H. Therefore H is an I-submodel of G. But H is not a proper I-submodel of G, for otherwise G and P do not satisfy the P-Minimality assumption. Hence G entails the exact same conditional independence statements as H does, which means that P is also Markov and faithful to G.

Therefore, for every probability distribution to which the CFA applies, any of the aforementioned weaker assumptions reduces underdetermination of causal structure by the distribution to the same extent as the CFA does.

Note that assumptions that are even weaker than the P-minimality assumption do not have this effect. In particular, if one adopts the causal Minimality assumption alone on top of the CMA (without also making the Triangle-Faithfulness assumption), the reduction of underdetermination is not nearly as great: $\mathbf{M}\text{-}\mathbf{M}(P)$ is typically much bigger than $\mathbf{M}\text{-}\mathbf{F}(P)$.[11]

We end this discussion with an open question. It is unclear whether it is the case that if $\mathbf{M}\text{-}\mathbf{AF}(P) \neq \varnothing$, then $\mathbf{M}\text{-}\mathbf{M}\text{-}\mathbf{TF}(P) = \mathbf{M}\text{-}\mathbf{AF}(P)$. In other words, we do not know whether for those distributions to which the Adjacency-Faithfulness assumption applies, it is always the case that the conjunction of the causal Minimality assumption and the Triangle-Faithfulness assumption reduces underdetermination to the same extent as the Adjacency-Faithfulness assumption does.[12]

In any case, it is fair to say that on the role of reducing underdetermination of causal structure by probability distribution, some weaker variations on the CFA actually do better. The extra strength of the CFA is to be seen in its other roles.

4 The computational role

Throughout this section, we assume that the input to causal discovery algorithms is a perfectly reliable oracle that can tell which conditional independence statements about $\mathbf{V}$ are satisfied and which are not by the probability distribution of $\mathbf{V}$. In practice the oracle is replaced by statistical tests of conditional independence based on samples from the distribution, and we will discuss some relevant issues in the next section.

[11] One way to see this is that the causal Minimality assumption, by itself, cannot rule out any causal order: for every ordering of the variables, there is a minimal DAG that satisfies the CMA.

[12] We do know that it is not true that whenever $\mathbf{M}\text{-}\mathbf{M}\text{-}\mathbf{TF}(P) \neq \varnothing$, $\mathbf{M}\text{-}\mathbf{PM}(P) = \mathbf{M}\text{-}\mathbf{M}\text{-}\mathbf{TF}(P)$. That is, there are cases in which the conjunction of the causal Minimality assumption and the Triangle-Faithfulness assumption reduces underdetermination to a greater extent than the P-Minimality assumption does. Raskutti and Uhler (2014) presented such an example in the proof of their Theorem 2.8. (b).

As we mentioned in Sect. 2, the facts stated in Propositions 1 and 2 suggest the following procedure:

SGS algorithm

S1. Form the complete undirected graph H on the given set of variables $\mathbf{V}$.
S2. For each pair of variables X and Y in $\mathbf{V}$, search for a subset $\mathbf{S}$ of $\mathbf{V}\backslash\{X, Y\}$ such that X and Y are independent conditional on $\mathbf{S}$. Remove the edge between X and Y in H if and only if such a *screening-off* set is found.
S3. Let K be the graph resulting from S2. For each unshielded triple $<X, Y, Z>$,
  (i) If X and Z are not independent conditional on any subset of $\mathbf{V}\backslash\{X, Y\}$ that contains Y, then orient the triple as a collider: $X \rightarrow Y \leftarrow Z$.
  (ii) If X and Z are not independent conditional on any subset of $\mathbf{V}\backslash\{X, Y\}$ that does not contain Y, then mark the triple as a non-collider.
S4. Execute some further orientation rules (the details of which do not matter here).

Basically, S2 is the step of inferring adjacencies and non-adjacencies, and S3 is the key step of inferring some arrow orientations. In light of Propositions 1 and 2, the two steps are obviously sound given the CMA and the CFA, but they are computationally expensive in that they almost always require a number of checks/tests of conditional independence that is exponential in the number of variables. Fortunately, the CFA allows both steps to speed up, as exemplified by the PC algorithm (Spirtes et al. 2000, pp. 84–85).

For S2, two strategies are employed in PC. First, for each of pair of variables X and Y, the search for a screening-off set for them is confined to their potential parents, i.e., subsets of other variables that are currently adjacent to X or subsets of other variables that are currently adjacent to Y. As adjacencies are pruned during the search, the number of conditioning sets that need be checked can be significantly reduced. Second, the search is done in stages, starting with the conditioning set of size 0 (i.e., the empty set), gradually increasing the size of conditioning sets, and stopping when the number of adjacent variables is smaller than the required size of conditioning sets. Call this procedure PC-S2.

For example, suppose we are given an oracle that is Markov and faithful to the causal structure in Fig. 1. In PC-S2, the procedure starts by checking, for each pair of variables, whether they are independent conditional on the empty set. The answer is no for every pair, so no adjacency is removed. Then it increases the size of conditioning sets to 1, and finds that $V_2 \perp\!\!\!\perp V_3|V_1$, $V_1 \perp\!\!\!\perp V_5|V_4$, $V_2 \perp\!\!\!\perp V_5|V_4$, and $V_3 \perp\!\!\!\perp V_5|V_4$. So the adjacencies between V_2 and V_3, between V_1 and V_5, between V_2 and V_5, and between V_3 and V_5 are removed. It then increases the size of conditioning sets to 2, and finds that $V_4 \perp\!\!\!\perp V_1|\{V_2, V_3\}$. So the adjacency between V_1 and V_4 is removed. Now the size of conditioning sets is increased to 3, but no more check is needed, because for every pair of the still adjacent variables, the number of other variables that are adjacent is fewer than 3. Hence, the search will not check any conditional independence statement with a conditioning set of size 3, in contrast to the SGS algorithm.

There are two main reasons why these strategies work in general. First, the CMA entails that two variables that are not adjacent in the causal DAG are either conditionally independent given one variable's parents or conditionally independent given the

other variable's parents. So if a screening-off set for two variables can be found, it can be found when the search is restricted to their parents. Second, the CFA, or the weaker Adjacency-Faithfulness assumption, entails that any conditional independence between two variables is sufficient to imply their non-adjacency. So at any stage, the true parents of a variable remain adjacent to that variable.

For S3, the PC algorithm simply checks, for each unshielded triple $<X, Y, Z>$, whether the screening-off set for X and Z found in PC-S2 contains Y or not. If the set does contain Y, the triple is inferred to be a non-collider. If the set does not contain Y, the triple is inferred to be a collider. Call this simplified step PC-S3.

To illustrate, in the previous example, consider the two (among several others) unshielded triples: $<V_2, V_1, V_3>$ and $<V_2, V_4, V_3>$. The PC algorithm simply recalls the screening-off set for V_2 and V_3 found in PC-S2, which is $\{V_1\}$. Since $\{V_1\}$ contains V_1, $<V_2, V_1, V_3>$ is inferred to be a non-collider. Since $\{V_1\}$ does not contain V_4, $<V_2, V_4, V_3>$ is inferred to be a collider. It is much more efficient than S3 in the SGS algorithm.

This significant simplification is justified by the CFA. For in view of Proposition 2 in Sect. 2, the CFA entails that either the antecedent of clause (i) in the original S3 or the antecedent of clause (ii) in the original S3 obtains. Since any conditional independence between X and Z is sufficient to either falsify the antecedent of (i) or falsify that of (ii), it is sufficient to either verify the antecedent of clause (ii) or verify that of clause (i).

In terms of bounds on the runtime complexity, a loose bound for the PC algorithm is $O(p^q)$, where p is the number of variables, and q is the maximal degrees (i.e., number of adjacent variables) of any variable in the true causal structure (Kalisch and Bühlmann 2007).[13] Thus for sufficiently sparse structures (i.e., with sufficiently small q), the PC algorithm is much more efficient than the SGS algorithm; the latter's complexity is almost always exponential in p.

Therefore, the CFA warrants a more efficient procedure than the original SGS procedure. This role, however, is not played as well by the weaker assumptions.

Consider first the Adjacency-Faithfulness assumption. It can still justify PC-S2. For that step only requires that any conditional independence between two variables imply non-adjacency, which is guaranteed by the Adjacency-Faithfulness assumption. However, it does not warrant PC-S3. The Adjacency-Faithfulness assumption does not entail Proposition 2 in Sect. 2. It leaves open the possibility that for an unshielded triple $<X, Y, Z>$, X and Z are both independent conditional on some set that contains Y and independent conditional on some set that does not contain Y. In other words, it is possible under the Adjacency-Faithfulness assumption that neither the antecedent of clause (i) in S3 nor the antecedent of clause (ii) in S3 is true. It follows that the simple check in PC-S3 is not sufficient.

One way to modify PC-S3 to make it work under the CMA and the Adjacency-Faithfulness assumption is presented in Ramsey et al. (2006). First, note that both clause (i) and clause (ii) of S3 are sound given the CMA alone. When the CFA is weakened to the Adjacency-Faithfulness assumption, we need to acknowledge the

[13] For the sample version of the PC algorithm, this bound holds with high probability.

possibility that neither clause (i) nor clause (ii) is applicable, in which case we should suspend judgment on whether the triple in question is a collider or a non-collider. Second, clause (i) and clause (ii) can still be improved under the Adjacency-Faithfulness assumption. Since the Adjacency-Faithfulness assumption implies that all the non-adjacencies resulting from PC-S2 are correct, we need only check, for each unshielded triple $<X, Y, Z>$, conditioning sets that are confined to subsets of X's potential parents (i.e., variables that are adjacent to X) and subsets of Z's potential parents (i.e., variables that are adjacent to Z). The rationale is essentially the same as that behind the similar strategy in PC-S2. The resulting procedure is called *Conservative PC* (CPC), on account of the possibility of it suspending judgment on whether an unshielded triple is a collider or not.

Consider again the example in which we are given an oracle Markov and faithful to the causal structure in Fig. 1. In CPC-S3, $<V_2, V_1, V_3>$ is not inferred to be a non-collider simply because the screening-off set for V_2 and V_3 found earlier, namely $\{V_1\}$, contains V_1. Rather, it has to also confirm that V_2 and V_3 are not independent given any subset of V_2's potential parents that does not contain V_1 or of V_3's potential parents that does not contain V_1. (In this case, those relevant subsets are $\varnothing$ and $\{V_4\}$.) We note, however, that despite the more involved CPC-S3, the overall complexity of the CPC algorithm is still bounded by $O(p^q)$, as is the PC algorithm.

More complications arise if the Adjacency-Faithfulness assumption is further weakened. Suppose we adopt the causal Minimality and Triangle-Faithfulness assumptions instead of the Adjacency-Faithfulness assumption. An immediate consequence is that the non-adjacencies inferred from S2 are not necessarily correct, for conditional independence between two variables does not entail non-adjacency if Adjacency-Faithfulness fails. We hasten to add, however, that the adjacencies inferred from the original S2 are still correct, which follows from the CMA alone. But if we use the more efficient PC-S2, then it is not even clear that the inferred adjacencies will be correct, for the current justification of PC-S2 depends crucially on the correctness of non-adjacencies. The worry is that if some adjacency is mistakenly removed, then some parent of some variable may be excluded in the search for a screening-off set for that variable and some other variable, which may lead to a false adjacency by failing to check some relevant conditioning sets. For these reasons, we suspect that PC-S2 is not always correct regarding adjacencies, though we have not found a concrete counterexample.

Exploiting the fact that S2 is correct on adjacencies (but may be incorrect on non-adjacencies), we proposed an algorithm named *Very Conservative SGS* (VCSGS) in Spirtes and Zhang (2014), which is provably correct under the CMA, causal Minimality assumption and the Triangle-Faithfulness assumption. The idea is that even though non-adjacencies resulting from S2 are only "apparent" (i.e., not necessarily correct), the original S3—with the added proviso that if neither the antecedent of (i) nor that of (ii) obtains, then the triple is marked as "ambiguous" or "unknown"—is still correct (by the CMA if the triple in question is really unshielded or by the Triangle-Faithfulness assumption if the triple in question is "apparently" unshielded but actually a triangle). In the end, the VCSGS algorithm checks whether every DAG compatible with the resulting (partial) graph is Markov to the input oracle. If yes, it confirms the non-adjacencies; if no, the non-adjacencies are marked as "ambiguous" or "unknown".

The exact proof of the correctness of the VCSGS is not important for our purpose here. The point is that VCSGS is more involved computationally than PC or CPC, or indeed SGS (though the worst case bound for VCSGS is on the order of that for SGS). Therefore, as we adopt increasingly weaker versions of the CFA, the computational costs of causal discovery increase significantly, at least for the constraint-based approach.[14]

5 The statistical role

The CFA entails that *any* conditional independence between two variables is sufficient to imply that they are not adjacent in the causal structure, and that, for an unshielded triple $<X, Y, Z>$, *any* conditional independence between X and Z is sufficient to decide whether it is a collider or a non-collider. These consequences imply not only that the number of conditional independence statements that need be checked can be much smaller than what is done in the SGS algorithm, but also that the search for a screening-off set for two variables X and Y need not go beyond the smallest size at which a screening-off set can be found. The PC algorithm, for example, always finds, for any X and Y that are not adjacent in the true causal graph, a minimal $\mathbf{Z}$ such that $X \perp\!\!\!\perp Y|\mathbf{Z}$ (in the sense that there is no set $\mathbf{Z'}$ such that $|\mathbf{Z}'| < |\mathbf{Z}|$ and $X \perp\!\!\!\perp Y|\mathbf{Z}'$, where $|\mathbf{Z}|$ denotes the number of variables in $\mathbf{Z}$). Moreover, the maximum size of a conditioning set checked by the PC algorithm is bounded above (with high probability) by the maximum degrees of any variable in the true causal graph, for by the time the algorithm is about to check conditioning sets of a size bigger than the maximal degrees, a screening-off set for each pair of non-adjacent variables will have been found, and so the adjacencies will have been sufficiently pruned that no variable is adjacent to sufficiently many variables to let the search continue.

The tendency to avoid checking conditional independence with a large conditioning set has important statistical advantages, for in practice, determining whether X and Y are independent conditional on $\mathbf{Z}$ depends upon performing statistical tests. When $|\mathbf{Z}|$ is large, the tests that are commonly used are statistically infeasible or inefficient for at least two reasons.

First, when the sample size is much smaller than the number of variables in $\mathbf{Z}$, standard statistical tests of conditional independence cannot be applied in a number of parametric families, including Gaussian, multinomial, and additive noise models. This is easiest to see in the multinomial case. $X \perp\!\!\!\perp Y|\mathbf{Z}$ entails that for every value $\mathbf{z}$ of $\mathbf{Z}$ (that has non-zero probability), $P(X = x, Y = y|\mathbf{Z} = \mathbf{z}) = P(X = x|\mathbf{Z} = \mathbf{z})P(Y = y|\mathbf{Z} = \mathbf{z})$. Standard statistical tests of $X \perp\!\!\!\perp Y|\mathbf{Z}$ estimate these three quantities and estimate the probability that the observed deviation from the equality is due to sampling error. It is possible to make maximum likelihood estimates of $P(X = x, Y = y|\mathbf{Z} = \mathbf{z})$, $P(X = x|\mathbf{Z} = \mathbf{z})$, and $P(Y = y|\mathbf{Z} = \mathbf{z})$ by counting the relative frequencies of $X = x$ and $Y = y$ for each value $\mathbf{z}$ of $\mathbf{Z}$. However, if $|\mathbf{Z}|$ is

[14] As we stress at the end, our discussion here is limited to the known constraint-based algorithms. We do not know whether similar points can be made for the score-based or the Bayesian approach, or even how weakening of the CFA bears on that approach. We thank an anonymous referee for raising this issue, and hope to explore it in future work.

large, then there are many possible values of **Z** even if all of the variables in **Z** are binary. For example, if there are 100 binary variables in **Z**, there are on the order of 10^{60} different possible values for **Z**. Even if the sample size is quite large, say on the order of 10^6, the vast majority of those possible values of **Z** will necessarily contain no sample points, and hence there is no unique maximum likelihood estimate of the relevant probabilities. Similarly, in the Gaussian case, a maximum likelihood estimate of the correlation of X and Y conditional on **Z** can be obtained by i) using the sample correlation matrix among X, Y, and **Z** to obtain a maximum likelihood estimate of the correlation matrix among X, Y, and **Z**, and ii) inverting the sample correlation matrix among X, Y, and **Z**. However, when the number of variables in X, Y, and **Z** exceeds the number of sample points, the sample correlation matrix is a singular matrix, and hence cannot be inverted to find an estimate of the partial correlation of X and Y conditional on **Z**. This is a practical issue in domains such as genetic regulatory networks, where there are thousands of variables, but only dozens of sample points: SGS is not applicable in these domains, whereas PC is.

Second, other things being equal, the power of the relevant statistical tests decreases as $|\mathbf{Z}|$ increases. In the Gaussian case, the power of a test increases with the "effective sample size", and the "effective sample size" is equal to the actual sample size minus $|\mathbf{Z}|$. In the Gaussian case, this entails that it is possible to perform statistical tests of conditional independence even for large $|\mathbf{Z}|$ (on the order of hundreds), as long as the sample size is also large (on the order of thousands), although rounding errors in the computation may make this practically infeasible. The problem is much more severe in the multinomial case, where the power of the standard (chi-squared) test is directly a function of $|\mathbf{Z}|$ (via the degrees of freedom of the chi-squared distribution), and when the sample sizes are in the thousands, tests of conditional independence with reasonable power are typically not feasible when $|\mathbf{Z}|$ is greater than 4 or 5.

In addition, we suspect that deviations from the parametric assumptions made by standard statistical tests have bigger effects on the reliability of the tests for larger $|\mathbf{Z}|$, but this requires more investigation.

We hasten to add that although the point here is related to the point on the computational role we made previously, both having to do with avoiding tests of conditional independence, they are *distinct* points. The computational point is about the *number* of tests, based mainly on the CFA's implication that *one* independence suffices to imply non-adjacency (and to imply orientations for unshielded triples); the statistical point is about the *order* of tests, based mainly on the CFA's implication that *any* independence (and so an independence with a small conditioning set, if any) suffices to imply non-adjacency.

To see it in another way, consider those algorithms that seek to first infer an undirected independence graph (a.k.a moralized graph), where there is an undirected edge between two variables if and only if they are independent conditional on all other variables, and then recover the DAG as much as possible from the undirected independence graph (Spirtes et al. 2000, pp. 124–125; Loh and Bühlmann 2013). When the undirected independence graph is sufficiently sparse, such a procedure can be computationally efficient, but thanks to the high-order conditional independence tests in the first step, can at the same time be statistically inefficient. Thus computational efficiency and statistical efficiency do not necessarily go together.

6 Conclusion

We have distinguished three perspectives to discuss the consequences of the CFA. From the perspective that philosophers are most sensitive to, that of reducing underdetermination of causal structure by probability distribution, the CFA is unnecessarily strong. However, its strength carries significant computational and statistical advantages.

Our discussion was limited to the known constraint-based algorithms in the literature. Whether similar points can be made for the score-based approach is an open question. Regarding computational complexity, a more ambitious project is to seek general results on the complexity of causal discovery under the CFA versus the complexity of causal discovery under the weaker assumptions (given certain bounds on the density of the causal structure).

Acknowledgments We thank two referees for helpful comments. The research of J. Zhang was supported in part by the Research Grants Council of Hong Kong under the General Research Fund LU342213.

References

Andersen, H. (2013). When to expect violations of causal faithfulness and why it matters. *Philosophy of Science Supplement*, *5*, 672–683.

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist*, *84*, 242–264.

Dawid, P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*, *41*, 1–31.

Hausman, D., & Woodward, J. (1999). Independence, invariance and the causal Markov condition. *British Journal for the Philosophy of Science*, *50*, 521–583.

Hitchcock, C. (2010). Probabilistic causation. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/causation-probabilistic/.

Hoover, K. D. (2001). *Causality in macroeconomics*. Cambridge: Cambridge University Press.

Hoyer, P., Janzing, D., Mooij, J., Peters, J., Scholkopf, B. (2008). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems* 21 (NIPS 2008), pp. 689–696.

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, *8*, 613–636.

Loh, P., & Bühlmann, P. (2013). High-dimensional learning of linear causal networks via inverse covariance estimation. arXiv:1311.3492v1 [stat.ML].

Pearl, J. (1988). *Probabilistic reasoning in intelligence systems*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge: Cambridge University Press.

Peters, J., Janzing, D., & Scholkopf, B. (2011). Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(12), 2436–2450.

Ramsey, J., Spirtes, P., & Zhang, J. (2006). Adjacency-faithfulness and conservative causal inference. *Proceedings of 22nd conference on uncertainty in artificial intelligence (UAI-06)*, pp. 401–408.

Raskutti, G., & Uhler, C. (2014). Learning directed acyclic graphs based on sparsest permutations. arXiv:1307.0366v3 [math.ST].

Shimizu, S., Hoyer, P., Hyvärinen, A., & Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*, 2003–2030.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Spirtes, P., & Zhang, J. (2014). A uniformly consistent estimator of causal effects under the k-triangle-faithfulness assumption. *Statistical Science*, *29*, 662–678.

Steel, D. (2006). Homogeneity, selection, and the faithfulness condition. *Minds and Machines*, *16*, 303–317.

Woodward, J. (1998). Causal independence and faithfulness. *Multivariate Behavioral Research*, *33*, 129–148.

Zhang, J., & Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference'. *Minds and Machines*, *18*(2), 239–271.

Zhang, J. (2013). A comparison of three Occam's Razors for Markovian causal models. *British Journal for the Philosophy of Science*, *64*(2), 423–448.

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2011). Kernel-based conditional Independence test and application in causal discovery. *Proceedings of the 27th conference on uncertainty in artificial intelligence (UAI-11)*, pp. 804–813.